1352.0.55.072

Australian Bureau of Statistics

Research Paper

A Review of Confidentiality Protections for Statistical Tables

With Special Reference to the Differencing Problem

Research Paper

A Review of Confidentiality Protections for Statistical Tables

With Special Reference to the Differencing Problem

Janice Wooton and Bruce Fraser

Statistical Services Branch

Methodology Advisory Committee

17 June 2005, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 16 FEB 2006

ABS Catalogue no. 1352.0.55.072
ISBN 0 642 48176 8

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Ms Janice Wooton, Statistical Services Branch on Canberra (02) 6252 5764 or email <janice.wooton@abs.gov.au>.

CONTENTS

A REVIEW OF CONFIDENTIALITY PROTECTIONS FOR STATISTICAL TABLES: WITH SPECIAL REFERENCE TO THE DIFFERENCING PROBLEM

Janice Wooton and Bruce Fraser
Australian Bureau of Statistics

1. CONFIDENTIALITY AND THE DIFFERENCING PROBLEM

The *Census and Statistics Act, 1905* provides the authority for the ABS to collect statistical information, and requires that statistical output shall not be published or disseminated in a manner that is likely to enable the identification of a particular person or organisation. This requirement means that the ABS must take care with any statistical information that relates to very small subpopulations or subsamples.

The techniques used to guard against identification or disclosure of confidential information in statistical tables are suppression of sensitive cells, and random adjustments to cells with very small values. These techniques have served the ABS well for standard statistical outputs that have been released through publication of statistical tables and other standard products. However, these methods, as currently implemented, are not effective for ad hoc output where a user can specify tailored tables. This may relate to a web-based (or other) service whereby a user can build a tailored table from a stored datacube or even direct from the survey microdata. These sorts of table-building services are commonly provided for users of Population Census data. Simple methods that may work well for a specified suite of standard tables may well be vulnerable to differencing of similar tables that are created from these sorts of table-building services.

An example of the differencing problem is where a user specifies a table for a user-defined geography, compiled from a number of small area building blocks. For example, the Statistical Local Area (SLA) "Remainder of ACT" had a population of approximately 430 at the time of the 2001 Census. The SLA consists of 7 Collection Districts (CDs) with populations of approximately 210, 115, 65, 25, 10 and two with populations of less than 5 each. If a user can specify tables for a tailored geography made up from CD building blocks, then they can specify a table for the full SLA, as well as a table for the amalgam of the six CDs with the greatest populations. Differencing the two tables provides information for a single CD with a population of less than 5 persons. The confidentiality protections applied to the SLA table, and the 6 CDs table, must therefore be sufficient to ensure that no information is disclosed through differencing the two tables. For further details regarding the definition of CD's and

SLA's see 1216.0 Australian Standard Geographical Classification (ASGC) – Electronic publication, July 2004.

The differencing problem is not limited to geography. User-specified tables could be requested for the whole population of "Remainder of ACT", and for the population aged less than 70 years, or for the population born in English-speaking countries, or for the population who only speak English at home. Differencing would again produce results in respect of small subpopulations.

In this paper we will review a number of techniques for protecting against identification or disclosure from tabular data, with special focus on the differencing problem. Sections 3 to 8 discuss 6 individual techniques for applying protections. We describe each method and discuss their advantages, disadvantages, and other characteristics. Section 9 summarises the discussion.

2. OUTLINE OF OPTIONS FOR ADDRESSING THE DIFFERENCING PROBLEM

The differencing problem could be addressed in one of the following ways.

1. By applying confidentiality techniques to the base microdata before producing tables or other statistical outputs. The most common techniques include data swapping, PRAM (Post-Randomisation Method), microaggregation, coding continuous data to categories (or categorical data to broader categories), introducing random perturbations, and deleting sensitive records. If sufficient confidentiality protections are applied directly to the microdata, then any tables or statistics produced using this data would be automatically protected. Tables and other statistical output would be consistent since disclosure control has been applied before generating the tables. However, the microdata confidentiality techniques introduce information loss which reduces the analytical value of the output.

2. By producing tables or statistical output from unprotected microdata and then applying protections to this output. The most common tabular techniques include cell suppression, collapsing rows and/or columns, and introducing random perturbations at the table cell level. These procedures are usually performed independently for each table/output or set of tables/output. This means that outputs may not be consistent. For example, if random perturbations are introduced at cell level then different perturbations could be applied to two independent requests for the same table. Independently applied protections can also allow disclosure, for example if two independent versions of the same table apply different cell suppressions.

3. By applying protections to both the microdata and the output, ie using both methods 1 and 2. If data swapping is applied to microdata, in many instances a cell perturbation method will also need to be applied to guarantee that differencing cannot occur especially in subpopulations and higher levels of geography which maybe unaffected by data swapping. For example, if we swapped a proportion of households between CDs in the same SLA then all cell totals at the SLA level would be unaffected by the swapping and the differencing problem could arise in respect of differencing by non-geographical variables.

The techniques canvassed below can each be assigned to either method 1 or method 2, i.e. they are applied to either the microdata or to the output. A final strategy, then, may well use a combination of the techniques canvassed below. However this paper does not propose or seek to discuss an overall strategy – rather the aim is to discuss and seek feedback on the application and characteristics of the individual techniques that can be employed.

Two further simplifications are made for the purposes of the discussion following. Firstly, only tabular output is considered, although tabular techniques can in general be extended, at least in theory, to other types of output, such as remote access to data

(see Schubert (2005) for further details). Secondly, the differencing problem is considered primarily in the context of geographical differencing and geographical output based on the 'building blocks' defined by the Australian Standard Geographical Classification (ASGC) – CDs, SLAs and higher levels of geography. The discussion can similarly be extended to other contexts of differencing by categorical classifications, e.g. age differencing using 'building blocks' of single year of age, five-year age ranges and other larger ranges; country of birth differencing based on the Standard Australian Classification of Countries; and so on. However, where relevant, the paper discusses the problems caused by protecting against all possible types of differencing.

It is also worth noting that this problem has been considered primarily in the context of 'frequency' data obtained from the Census of Population and Housing and household surveys. This is in contrast to 'magnitude' data that is produced for business surveys, where a single contributor can dominate a cell total. While many of the techniques are applicable to magnitude data, the special problems introduced by it are not considered in this paper.

3. TECHNIQUE 1: INDEPENDENT PROTECTION OF COMPONENT 'BUILDING BLOCKS'

One method for dealing with geographical differencing originates from the UK, although it has never been formally evaluated or implemented anywhere. The method is as follows. Each table for a non-standard geographic area will be defined by the user as an amalgam of standard building blocks. The definition would utilise all levels of the standard areas (i.e. as defined in the ASGC), from the largest geographic area (State) to the smallest (CD). For example, a particular area may be defined as the sum of one SLA and four CDs. Tables for each of these standard component areas would be produced, the tabular protections (e.g. cell perturbation) would be applied to each component table, and the tables would be summed to give the final confidentialised result. This means, for the example of an area given by one SLA and four CDs, that five component tables would be produced and independently protected, then combined to produce the final table. An area of interest may also be defined in terms of standard areas via subtraction e.g. one SLA minus two CDs. Here, component tables are created and protected for the SLA and for each of the two CDs, and then each CD table is subtracted from the SLA table.

If this method is used, it could be decided that protection action need only be taken for a subset of 'building block' tables. For example, it may be decided that SLA level tables have a large enough population to make protection action unnecessary. Then protection would only need to be applied to component CD tables. In practice, it is more likely that a population threshold will be specified, e.g. no protection action required for tables that report on a population of size 300 or greater. It may also be necessary to take the number of cells into account, e.g. base a decision on the average population per table cell.

This method would result in more information loss the more a user-defined area differed from a standard area. The method can be adapted to work for averages and ratios (by including numerator and denominator in the table specification, and deriving the final table ratio from the final table numerator and denominator) but separate solutions would be needed for percentiles and indexes.

The following characteristics are noted:

1. If table subtraction were allowed (e.g. an SLA minus a CD) in conjunction with cell perturbation protection, then the final table could contain negative values. If these negative differences were then rounded to zero a bias would be introduced. It may therefore be desirable to stipulate that user defined areas can only be specified as sums of standard building blocks – although this may lead to a greater level of information loss.

2. Complex table specifications that are open to potential differencing across a number of variables (e.g. user defined area for 'baby boomers' born prior to 1965 born in English-speaking countries) result in complex specifications for component tables, and high information loss due to the protection of a high number of small population tables.

3. With this method we are essentially assuming that the standard areas are the key output areas for which we want to minimise information loss. Any user defined table which differs greatly from standard areas is in general subject to greater information loss to protect against disclosure through differencing. Key users may be disadvantaged by this. For example, a CD is designed for collection purposes so that interviewers can collect the required information in an efficient way – it is not designed to meet user needs for analysis or output, and consequently are not optimal for certain types (if any types) of analyses. On the other hand, environmental catchment areas or electoral division areas can be important for certain types of analysis or output. Should we place so much emphasis on standard areas when many users may use non-standard areas and have important reasons to want their results to be as accurate as the standard area tables? {Note – the current CD is being phased out for output purposes in recognition of its design as a collection geography, not an output geography}

4. TECHNIQUE 2: CELL PERTURBATION

Cell perturbation is a technique which confidentialises tables by introducing perturbations to interior cell counts. The technique protects against all types of differencing (geographical and subpopulation differencing). An advantage of the method is that information loss and disclose risk are measurable, but a disadvantage is that if perturbed marginals are generated by adding the perturbed cell values (to retain additivity), then the information loss on the marginals is high. A particular technique is illustrated below. The technique works in a similar way to the random rounding base 3 method where 1's and 2's are rounded to 0 or 3.

Denote the i^{th} interior cell count of a multi-way table as n_i. For each non-zero n_i simulate a set of independent and identically distributed deviations d_{ij} for $j = 1, 2, \ldots, n_i$ according to the probability distribution:

$$P(d_{ij} = 0) = p_i,$$
$$P(d_{ij} = -1) = \frac{2}{3}(1 - p_i),$$
$$P(d_{ij} = 2) = \frac{1}{3}(1 - p_i),$$

where $p_i = \begin{cases} 0 & \text{if } n_i \le x \\ 1 - \dfrac{x}{n_i} & \text{otherwise,} \end{cases}$

And x is some positive number. x is set so that any cell frequency less than or equal to x is perturbed with certainty. For cells with frequency greater than x, x is the expected number of non-zero d_{ij} deviations added to the cell.

The choice of a distribution for d_{ij} here has been arbitrary, and influenced by the rounding to base 3 methodology. Note that a cell frequency of 1 can only be perturbed to a value of either 0 or 3. Similarly, if x is set to be 2 or greater, then a cell frequency of 2 can only be perturbed to a value of 0, 3 or 6. This means that if a cell frequency of 1 or 2 is observed following the perturbation, then it is certain that the true (i.e. unperturbed) cell frequency is not a 1 or a 2 (if $x \ge 2$). The method can be fine-tuned by looking at different distributions of d_{ij} (or different distributions for the sum of the d_{ij} perturbations contributing to a cell).

Once all the d_{ij} have been simulated we then re-calculate the ith interior cell count as $n_i^* = n_i + \sum_{j=1}^{n_i} d_{ij}$. The marginal and overall totals are then re-calculated by adding the relevant n_i^* values together. We have now formed the adjusted table.

Note: Here we are assuming $n_i \geq 1$. If $n_i = 0$ then let $n_i^* = n_i$ and we have no cell adjustment and structural zeros are maintained.

The logic behind this method lies in perturbing each cell to introduce enough variability to ensure that any differencing resulting in small numbers is sufficiently protected. Cell frequencies of 1 or 2 are perturbed with certainty (as long as x has a value of 2 or greater), and other small cell frequencies of 3 or greater are perturbed with high probability.

Expected Value and Variance of n_i^* (assuming that $n_i \geq 1$)

$$\begin{aligned} E(n_i^*) &= n_i + \sum_{j=1}^{n_i} E(d_{ij}) \\ &= n_i + \sum_{j=1}^{n_i} (-\frac{2}{3} p_i + \frac{2}{3} p_i) \\ &= n_i. \end{aligned}$$

Therefore the adjusted interior cell counts are unbiased (it also follows that marginal and overall totals are unbiased since to obtain these we are adding together unbiased terms).

$$\begin{aligned} Var(n_i^*) &= Var(n_i + \sum_{j=1}^{n_i} d_{ij}) \\ &= \sum_{j=1}^{n_i} Var(d_{ij}) \\ &= \sum_{j=1}^{n_i} \frac{2}{3}(1-p_i) + \frac{4}{3}(1-p_i) \\ &= 2n_i(1-p_i) \\ &= \begin{cases} 2n_i & \text{if } \; n_i \leq x \\ 2x & \text{otherwise} \end{cases} \end{aligned}$$

What if a cell is obtained by differencing, e.g. a cell frequency of n_j^* differenced from a cell frequency of n_i^* (with $n_i^* \geq n_j^*$)?

$$E(n_i^* - n_j^*) = n_i - n_j,$$

$$Var(n_i^* - n_j^*) = 2n_i(1-p_i) + 2n_j(1-p_j).$$

That is, the differenced table is subject to greater variability within each cell than if it had been generated and perturbed independently.

Information Loss Versus Disclosure Risk and our choice of x

Our choice of x will allow us to control the balance between information loss and disclosure risk. Obviously as information loss increases disclosure risk decreases. We need to be able to come up with a compromise between these two conflicting constraints to obtain an optimal choice of x.

Because our cell adjustment method is unbiased the loss of information for the i^{th} cell can be measured using $Var(n_i^*)$ and the maximum value of this variance term will be $2x$. Also, because the cells are randomly adjusted independently of one another, it follows that the variance of marginal and overall totals can be calculated as the sum of the variances of the interior cells which add to give them. For example, suppose a multidimensional table had k interior cells, then the variance of the grand total would be at most $2xk$. Based on these measures of information loss we can come up with a maximum allowable value of x (if x is too large then the tables would be useless for analysis). If the variance of the various marginal and overall totals were too high we could actually apply the cell randomisation rule separately to the cells containing the totals and they would then have a variance of $2x$ instead of $2xk$, but then the tables would not have the additivity property.

Disclosure risk is a little bit harder to measure than information loss. If a user only had access to one table, then the disclosure risk would be minimal. In this case if $x \geq 2$, then n_i^* can only be equal to 1 or 2 if $n_i \geq 3$ and all population uniques (i.e. cells with n_i equal to 1 or 2) would be protected from confidentiality breaches. Suppose now that a user had access to two tables, where the second table differed only slightly from the first. For example, the second table might only differ by one CD in definition. It would then be possible for the user to subtract the two tables and obtain detailed information for the one CD. Because we have introduced variability into the interior cell counts in both tables, the user would not be sure that an observed difference in the cell counts of 1 or 2 was really a difference of 1 or 2. To minimise

disclosure risk we would like the variance of the difference in corresponding observed cells from similar tables to be large enough to ensure that the probability of an observed difference of 1 or 2 is small given that the actual difference is 1 or 2.

As a worse case scenario, suppose that we had two tables (Table 1 and Table 2) representing two different subpopulations where the second subpopulation is contained within the first. A third subpopulation table (Table 3) could be calculated by subtracting the second table from the first. Now suppose that the cell count in the r^{th} cell in Table 1 is $n_{r,1}$ and the cell count in the r^{th} cell in Table 2 was $n_{r,2}$ where $n_{r,1} - n_{r,2} = 1$. We need to be able to introduce enough randomisation into the confidentialised tables to ensure that $P(n^*_{r,1} - n^*_{r,2} = 1)$ is small enough to ensure we would not have a sufficiently large disclosure risk. We will now investigate this hypothetical worse case scenario by simulation to compare a few different choices of x.

Empirical Probabilities of obtaining a difference of 1 given a difference of 1 (disclosure risk)

The following table contains the empirical probabilities of obtaining a difference of 1 given that $n_{r,1} - n_{r,2} = 1$, for different values of $n_{r,1}$ and x. Note that the empirical probabilities are calculated using a sample size of 10,000 and the probabilities are therefore approximate.

$n_{r,1}$	x	$P(n^*_{r,1} - n^*_{r,2} = 1 \mid n_{r,1} - n_{r,2} = 1)$	$P(n^*_{r,1} - n^*_{r,2} = 1 \text{ or } 2 \mid n_{r,1} - n_{r,2} = 1)$
1	2	0	0
2	2	0	0
3	2	0.18	0.28
4	2	0.14	0.28
5	2	0.14	0.28
5	4	0.12	0.18
7	6	0.10	0.15
20	2	0.15	0.28
20	4	0.10	0.20
20	6	0.08	0.16
100	2	0.15	0.29
100	4	0.10	0.20
100	6	0.09	0.17
1000	2	0.15	0.29
1000	4	0.11	0.21
1000	6	0.09	0.16

Based on the above table we can now make a decision about our choice of x. We need to decide on our minimum allowable disclosure risk and choose the smallest

value of x which satisfies this constraint. A likely candidate for our choice of x is $x = 2$ (unless of course it is decided that the disclosure risk is too large for this choice). To get an understanding of the range of $n^*_{r,1} - n^*_{r,2}$ the probability distributions of $n^*_{r,1} - n^*_{r,2}$ can be investigated. As an example, with $x = 2$ and $n_{r,1} = 5$ we get the following empirical distribution of $n^*_{r,1} - n^*_{r,2}$:

Observed difference	*Frequency*	*Empirical probability*
–9	4	0.0004
–8	11	0.0011
–7	26	0.0026
–6	80	0.0080
–5	146	0.0146
–4	296	0.0296
–3	521	0.0521
–2	777	0.0777
–1	1,145	0.1145
0	1,302	0.1302
1	1,411	0.1411
2	1,434	0.1434
3	1,030	0.1030
4	766	0.0766
5	512	0.0512
6	304	0.0304
7	148	0.0148
8	55	0.0055
9	20	0.0020
10	5	0.0005
11	5	0.0005
13	2	0.0002

Clearly from the above information the user cannot have much confidence in the real value of any small observed differences.

Protection against the disclosure occurring through 100% cells

The 100% disclosure problem is illustrated in the table below. It occurs when a row or a column contains only one non-zero cell. 100% of the row/column value is found in a single cell, and in the example below, allows the user to deduce that all males selected in the survey are aged 20–29 years. If any male sample selection has been selected with certainty, then it follows that this person must be aged 20–29 years. In a census situation the full population has been selected with certainty, and the table discloses that all males are aged 20–29 years. Note that this table also discloses that anyone aged 0–19 is female.

	0–19	*20–29*	*30–49*	*50+*	*All ages*
Male	0	3	0	0	3
Female	3	3	0	0	6
Persons	3	6	0	0	9

The cell perturbation method we have discussed above provides some protection against disclosure from 100% cells because $n_i^* = 0$ does not imply $n_i = 0$. For example, a user who has been provided with the above table cannot be sure that all of the observed 0's are true zeros since it is possible that some cells were originally 1's or 2's, in which case there is a high probability that these were rounded to zero. It is also possible that the original cell counts could have been 3, 4, 5, …, etc. (with decreasing probability as n_i increases).

Summary: Advantages of the method

1. Relatively easy to implement.
2. Information loss versus disclosure risk is measurable. If $x = 2$, variance of an interior cell total is at most 4. This is not a large variance and we do not expect the counts n_i to be completely accurate anyway due to non-sampling error.
3. The larger the cell count n_i, the smaller the relative perturbation error of n_i^*.
4. Relative perturbation error of n_i is high for small cell counts (this is good because it will give added protection for small cell counts where the risk of disclosure is higher).
5. Protects against geographical and non-geographical differencing.
6. All user defined areas or subpopulations can be output (subject to the number of cells being less than the population count).
7. Provides some protection against disclosure occurring through 100% cells.
8. The method is unbiased.
9. *additivity* i.e. interior cells add to totals

Summary: Disadvantages

1. consistency i.e. the same table cells or totals are not consistent across tables since each table is randomly perturbed independent of one another.
2. Information loss is relatively high on the table marginals, especially for tables with a large number of interior cells. This disadvantage may be addressed by modifying the basic method, e.g. allowing p_i to vary with k, the number of interior cells in a table, as well as x_i and n_i. However, this would mean increasing the disclosure risk for tables with large numbers of interior cells.

5. TECHNIQUE 3: RANDOM ROUNDING

Random rounding implements rounding to base k. Common values of k are 3 (which can be thought of as the minimum acceptable value for k) and 5 (which offers greater protection for greater information loss, and produces output where the rounding is more readily apparent – multiples of 5 are more readily identified than multiples of 3).

The rounding is performed independently for each cell. Additive tables can then be produced by adding together the perturbed interior cells, with the result that the net perturbation applied to the marginals is relatively high, and increases as the number of interior cells in the table increases. Alternatively, the marginals can be independently rounded to base k, with the result that the final table is in general not additive (the sum of the entries may not be equal to the totals in the marginals). Random rounding addresses both the geographic and subpopulation differencing problems.

Statistics Canada use a random rounding to base 5 for Population Census tables. Marginals are rounded independently of interior cells (resulting in loss of additivity), and rounding is performed independently each time a table is produced (resulting in loss of consistency). Differencing of rounded tables will therefore result in differences that are a multiple of 5, e.g. –5, 0 or 5. An intruder cannot infer the true value from the rounded difference. However, if the same table requests are made many times over, and averages are taken of the independent rounding in each cell, then the law of large numbers implies that the averages will converge to the true frequencies (because the rounding algorithm used is unbiased). However, the number of trials required to get reasonable precision is in the seventies. Protections are required to ensure that no client can request such a large number of repeated requests.

Let us now compare the rounding to a base method to the cell randomisation method that was introduced earlier.

For simplicity, suppose we are randomly rounding to base 3. This means that cell counts that are not multiples of 3 get rounded randomly up to the nearest multiple of 3 or randomly down to the nearest multiple of 3. Let n_i denote the original cell count of the i^{th} cell and let n_i^* denote the rounded cell count where $n_i^* = n_i + d_i$ and d_i is a random deviation equal to either 0, 1, 2, –2 or –1.

All the d_i are generated independently with the following distribution:

$$P(d_i = 0) = \begin{cases} 1 & \text{if } n_i = 0 (\text{mod } 3) \\ 0 & \text{otherwise} \end{cases}$$

$$P(d_i = 1) = \begin{cases} \frac{2}{3} & \text{if } n_i = 2 (\text{mod } 3) \\ 0 & \text{otherwise} \end{cases}$$

$$P(d_i = 2) = \begin{cases} \frac{1}{3} & \text{if } n_i = 1 (\text{mod } 3) \\ 0 & \text{otherwise} \end{cases}$$

$$P(d_i = -1) = \begin{cases} \frac{2}{3} & \text{if } n_i = 1 (\text{mod } 3) \\ 0 & \text{otherwise} \end{cases}$$

$$P(d_i = -2) = \begin{cases} \frac{1}{3} & \text{if } n_i = 2 (\text{mod } 3) \\ 0 & \text{otherwise} \end{cases}$$

Given the above distribution we can work out the variance of n_i^*.

$$\begin{aligned} \text{var}(n_i^*) &= \text{var}(d_i) \\ &= \begin{cases} 0 & \text{if } n_i = 0 (\text{mod } 3) \\ 2 & \text{otherwise} \end{cases} \end{aligned}$$

and note that the method is unbiased (this is easily proved using the distribution of d_i).

Note: If a rounding base of 5 had of been used then,

$$\begin{aligned} \text{var}(n_i^*) &= \text{var}(d_i) \\ &= \begin{cases} 0 & \text{if } n_i = 0 (\text{mod } 5) \\ 4 & \text{otherwise} \end{cases} \end{aligned}$$

and the variance function is similar to what was obtained in the cell randomisation technique discussed previously.

Comparison of techniques 2 (cell perturbation) and 3 (random rounding)

The main differences in the two techniques are:

1. With cell perturbation, only zeros are guaranteed to remain unchanged by the protection, whereas with random rounding, any multiple of base k will remain unchanged.
2. With cell perturbation, the resulting rounded table looks more 'natural' since the counts are not all multiples of a rounding base. With random rounding it is more obvious that a rounding protection has been applied.
3. There are differences in the variance functions, which allows the different levels of information loss to be compared for the two methods.

The information loss associated with the two methods can be quantified as follows. Entropy or information entropy of a random variable X is the expected quantity of information given by knowledge of the outcome of X. It is denoted by $H(X)$ and is given by the formula:

$$H(X) = E(-\log_q(p(X))) = -\sum_{x \in X} p(x)\log_q(p(x)),$$

where $p(X)$ is the probability density function of X, q is some integer and the summation is over all possible values of X.

Conditional Entropy is the expected additional information given by knowledge of X, if Y is already known and is denoted by $H(X|Y)$ and is given by the formula:

$$H(X \mid Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x \mid y)\log_q(p(x \mid y)).$$

$H(X|Y)$ measures the ambiguity of X after knowing Y.

Let the random variable X denote our original cell count and for simplicity (and because we are mostly concerned about protecting against disclosure of small cell counts) assume that X can only take on the values 0, 1, 2, 3, 4 or 5. Let Y denote the observed cell count after the cell has been randomised. We can measure the expected information gain about X (the original cell counts) after observing Y (the randomised cell counts) by using the following formula:

$$\frac{H(X) - H(X \mid Y)}{H(X)} * 100\%.$$

We now compare the random rounding to base 5 technique with the cell perturbation technique using $x = 2$. In order to do this though, we need a prior probability distribution for X. Note that $H(X)$ measures the uncertainty about X before we observe X. If we are very uncertain about the outcome of X, then we will obtain a

large amount of information by observing X and this implies that $H(X)$ is large. A distribution for X which maximises $H(X)$ is $p(X)=1/6$ for $X = 0, 1, 2, 3, 4, 5, 6$. Assuming that we have no prior knowledge about the distribution of X, then this distribution is the best prior distribution that we can use for X. Because there are 6 possible outcomes it is best to let $q=6$, because this implies that $H(X)=1$ and we have standardised the entropy. With the above in mind we have calculated the expected information gain for both the methods. All calculations were done in SAS and $p(y)$ and $p(x|y)$ were calculated via simulation based on 1 million simulated values for X and Y. The summarised results are:

For random rounding to base 5:

$$\frac{H(X)-H(X|Y)}{H(X)}*100\% \approx 17\%.$$

For cell perturbation with $x = 2$:

$$\frac{H(X)-H(X|Y)}{H(X)}*100\% \approx 33\%.$$

Therefore, on average the cell perturbation technique will give us more information about the original cell counts than random rounding to base 5 and there has been less information loss.

If we now assume that the distribution of X decreases slightly as X increases (which could occur in practise), for example:

X	*0*	*1*	*2*	*3*	*4*	*5*
p(X)	0.35	0.25	0.15	0.10	0.10	0.05

then we obtain the following similar results:

For random rounding to base 5:

$$\frac{H(X)-H(X|Y)}{H(X)}*100\% \approx 17\%.$$

For cell perturbation:

$$\frac{H(X)-H(X|Y)}{H(X)}*100\% \approx 34\%.$$

Controlled random rounding

When we apply a rounding method to the interior cells in a table, in order to calculate the marginal and overall totals we would like to add up the relevant interior rounded cells to obtain the totals. This method often leads to totals that might be too 'variable'. If we round the cell totals separately we can then lower the variance, but the resulting table is no longer additive. To solve this problem the ONS (Office of National Statistics) in the UK use controlled random rounding on some of their tables.

As discussed by Lowthian and Merola (2004), the idea of controlled random rounding was proposed a long time ago, but only recently a program has been developed to round in a controlled manner large multidimensional and linked tables. Lowthian and Merola (2004) give a brief description of the controlled random rounding algorithm, then discuss some of its disclosure control properties and how the algorithm is embedded in the software package τ-Argus. The paper gives reference to Salazar–Gonzalez (2002) and Salazar–Gonzalez *et al.* (2004) who describe both controlled rounding theory and practise in more detail. In summary an algorithm is implemented which compromises between rounding interior cells and rounding various marginal/totals to a particular base in a more optimal way, ensuring that the tables are still additive.

However, as mentioned by Shlomo (2005), controlled random rounding still needs some development work to deal efficiently with census sized tables.

6. TECHNIQUE 4: DATA SWAPPING/PRAM (POST-RANDOMISATION METHOD)

The main idea behind data swapping/PRAM in this context is to introduce uncertainty to the geographic location of households in the census data set, or similarly uncertainty to other characteristics. In the following we will briefly outline a few specific ways in which data swapping/PRAM could be applied.

Simple geographic transitions

The geographical differencing problem has mainly arisen due to increased demands for small area statistics from the Census of Population and Housing as well as from administrative datasets and other sources. Small area statistics are a confidentiality risk because people usually know more about their neighbours than people living further. This problem can be addressed by swapping a proportion of households from one geographical area to another. The swap can either be localised, e.g. between two CDs in the one SLA, or more extreme, e.g. swapping households between States. The advantage of a localised swap is that there is no information loss at the SLA level, but the disadvantage is that it offers less protection due to the smaller level of perturbation. This can be addressed by swapping at a number of levels, e.g. identify records that are rare within their CD and swap them within their SLA, then identify records that are rare within their SLA and swap them between States.

The Post-Randomisation Method is an implementation of the data swapping idea whereby every record has some probability (usually small) of being swapped. We illustrate through an example.

Suppose a SLA contains 4 CDs m_1, m_2, m_3 and m_4 with x_1 households in m_1, x_2 households in m_2, and so on.

Then according to the following markov transition matrix where the rows are the initial state and the columns are the final state we can randomly change the location of each household from its initial CD to a new CD in the same SLA with transition probability specified by the following Markov transition matrix (assuming that $x_1 \leq x_2, x_3 \; and \; x_4$):

$$
\begin{array}{c|cccc}
 & m_1 & m_2 & m_3 & m_4 \\
m_1 & 1-p & \frac{p}{3} & \frac{p}{3} & \frac{p}{3} \\
m_2 & \frac{px_1}{3x_2} & 1-\frac{px_1}{x_2} & \frac{px_1}{3x_2} & \frac{px_1}{3x_2} \\
m_3 & \frac{px_1}{3x_3} & \frac{px_1}{3x_3} & 1-\frac{px_1}{x_3} & 1-\frac{px_1}{x_3} \\
m_4 & \frac{px_1}{3x_4} & \frac{px_1}{3x_4} & \frac{px_1}{3x_4} & 1-\frac{px_1}{x_4}
\end{array}
$$

where $p \in (0,1)$.

By applying this markov matrix to each of the households in the SLA we have added some uncertainty (measured by p) into CD-level tables. If p is small then the level of information loss is small – and further can be accurately described to the table users by releasing the transition matrix. It can be stated that no user can be sure that any apparent disclosure is accurate, since they cannot be sure whether the identified record has been swapped. However, if p is small a user can still be confident that an apparent disclosure is accurate. Determining a value of p that provides adequate protection is difficult (although one could make use of expectatation ratios to provide some guidance as described in Gouweleeuw *et al.* (1998)). The main advantage of this method is that it allows for a detailed description of the method of perturbation, and hence a quantification of information loss that can be used by analysts, without unduly jeopardising the protection provided by the method.

Another characteristic of the method is that it provides some protection against both geographical differencing and subpopulation differencing, even if records are only moved geographically. An observed cell count of 1 in the differenced table does not necessarily imply that the real cell count is 1 since households containing units in the relevant category may have been swapped from a different geographical area. The form of the transition matrix above also implies that on average after PRAM has been applied there will still be x_1 households in m_1, x_2 households in m_2, and so on. Exact counts can be guaranteed by implementing data swapping instead. That is, for each household that moves out of a CD there is one that comes back in to replace it.

The transition matrix above will only move a household within an SLA, ie it will only slightly perturb location. Because of this, we are less likely to distort some location dependent variables' distributions. However, if clusters of similar households occur within CDs, by swapping or PRAMing them we are making the distributions more heterogeneous at the CD level, which is not an advantage. Because the data swapping

is done at a low geographic level, we cannot guarantee that even person counts (as opposed to household counts) will remain unaffected. For example, if household size is related to CD, as can occur where some CDs contain mainly smaller apartments while other CDs contained mainly houses. Because the data swapping is done at such a low geographic level we have not protected against potential differencing that could occur at higher geographic levels, especially if subpopulation tables are of high dimension with respect to demographic variables. This weakness can be addressed by applying a second level of swapping / PRAMing, eg swapping households among SLAs within a State. A similar transition matrix can be used, and may well use a smaller value of p for inter-SLA transitions than for the intra-SLA transitions.

Constrained geographic transitions (take other variables into account when data swapping)

If the number of households is large enough in a particular geographic area, then we can begin to take other variables into account when data swapping. This will ensure that some key distributions of variables can be maintained below the level of geography where swapping takes place and further protect against geographical differencing that might occur at higher geographical levels. In the previous section we focussed on swaps of households between CDs within a SLA. If we wanted to introduce restrictions to preserve the distributions of other variables then it may no longer be viable to do within SLA swaps – we may need to swap within larger geographies to ensure suitable swapping partners can be found. The swapping rule would stipulate that only households with similar characteristics could be swapped. For example, a swapping rule might be that we only swap with households with the same number of people, in the same broad age by sex groups.

As the geographical region where we allow swaps to take place increases, we can control for more auxiliary variables when swapping the locations of the households. However, there are a few problems with this. Suppose the dataset has n variables and k of these have been used to form the swapping groups. Essentially what we are doing is keeping the first k variables unchanged in distribution and the relationships between these remain the same. But because the other $n–k$ variables have been swapped, the relationship between any of these with the other k variables will be distorted at the various geographic levels below which swaps have been made. How does one decide which variables to use when forming swapping groups? And how does one appropriately balance the number of variables k and the size of the geographical region within which swaps are made?

Overseas experience with Population Census data swapping

One example of where data-swapping has been applied to protect against geographical disclosure is in the US Population Census. They use a long and a short form, with only the short form distributed to the entire population. The short form only collects a small number of data items. Their method involves swapping households with similar demographic variables across any level of geography where swapping partners can be found. This method introduces enough uncertainty about the location of households such that they can be reasonably sure the geographical differencing problem is solved for their tables at all levels. In their publications they do not use rounding (The US Bureau of Census must be confident that data swapping has in addition introduced enough uncertainty that it also protects against subpopulation differencing as well). Rounding of cells is only used on special tabulations that users want access to (this ensures that non-standard tables are sufficiently protected against differencing).

Note that the much larger U.S. population makes data-swapping more feasible and easier to implement than in Australia – it is easier to find swapping partners that match on a large number of variables.

Currently research is being done in the UK on data swapping. The fact that there is still being research done on the method indicates that data-swapping is very hard to implement 'optimally'. The effects on distributions and statistical analyses can be very hard to measure.

Summary

Data swapping and PRAM are methods applied to microdata before tabulations and other statistical output are created. For this reason, the methods guarantee consistency of output. This characteristic, and the fact that the confidentiality protections are applied 'up front' before any output is generated rather than as an additional step to go through to produce output, are the main advantages of the methods. The main disadvantages are listed below.

1. The methods only partially solve the differencing problem. It depends how one forms the criteria through which households are swapped/PRAMed. This is why data swapping is often used in conjunction with rounding or perturbation of cell values to ensure that differencing of subpopulations can not occur (e.g. US Bureau of Census special tables methodology). Combining the two methods also provides two layers of protection against geographic differencing. However cell rounding or perturbation can lead to inconsistent output, and needs to be performed at the time the output is generated, negating the main advantages of the swapping/PRAM methods.

2. The basic PRAM methodology we have described does not guarantee that identifiable records will be perturbed. A data swapping method can be designed so that identifiable records are identified and then swapped with certainty, however this adds complexity and cost to the method. PRAM and swapping rely to some extent on the users uncertainty as to whether or not the data has been perturbed. A worst case scenario for a statistical agency relying on these methods may be where a user makes a disclosure, publicises it, and the agency then discovers that no perturbation was applied to this record. It would be difficult to base an effective defence on the fact that there was some non-zero probability that a perturbation would have been applied. It is perhaps the case that the main protection provided by swapping/PRAM is to reduce the motivation of an intruder to attempt identification, and that additional protections such as cell rounding / perturbation is also required to provide protection against intrusions that are attempted.

3. It is difficult to fully quantify disclosure risk and information loss with these methods, although PRAM transition matrices and methods can be published to help analysts account for the information loss that is introduced. See Gouweleeuw *et al.* (1998) for further details. A discussion of how to incorporate PRAM transition matrices into multivariate analyses is given in Van Den Hout and Van Der Heijden (2004).

7. TECHNIQUE 5: MICRODATA PERTURBATION THROUGH ESTIMATION WEIGHTS

One desirable property to build into a methodology is to ensure that a particular estimate is released in the same way in any table that contains the estimate. Many methods perturb or round cells independently between tables, and so a particular estimate may be rounded or perturbed upward in one table, and downward in another. One way to build in consistency between tables is to use a sampling method to round.

Consistency is maintained when disclosure control is applied to the microdata before tabulations are made. Suppose we took a without replacement simple random sample of census households with a sampling fraction of 1/3 say. When forming all the tables we can only use the 1/3 of households that were chosen in the sample and multiply all cell counts by 3. All tables would then be consistent and we have protected ourselves against differencing and the 100% cell issue. Although this method ensures consistency, it is not ideal because we have lost 2/3 of the information contained in the sample! Also, population uniques are not sufficiently protected since we know that an observed cell count of 3 or 6 is a sample unique and could quite possibly be a population unique. However, we can improve the above sample design.

By using household size (number of people within a household) and a small geographic area such as CD as stratification variables we can improve on the above sample design. It seems reasonable to expect that household size and geography is correlated with many individual level and household level data items. The differencing problem can occur in many different subpopulations, but protecting against geographical differencing is seen as the most important issue. With this in mind we suggest the following sample design:

1. Place each household in the microdata file into a strata defined by CD and household size groups 1, 2, 3, 4, 5+.
2. Within each strata we perturb the record weight by a multiplicative factor, referred to as the perturbation weight. The perturbation weight that is applied to a household is also used for all persons within a household. If we are working with a census dataset then the initial weights are 1, but are multiplied by the perturbation weight to produce a final weight that can take values other than 1. Perturbation weights of 0 are permissible and will lead to final record weights of 0 – such a weight means that the household and the people within the household to not contribute to final estimates (analogous to the household not being sampled).
3. Tables are produced using weighted aggregates. If the final weight is 0, the unit does not contribute to any cell count. If the final weight is 2 and the unit falls in a particular cell then it contributes 2 to that cell etc.

4. The perturbation weights can be assigned to each household independently of each other within each non-empty strata using the following probability distribution:

$$P(w_{ik} = 0) = \frac{3}{5a_k}$$

$$P(w_{ik} = 1) = 1 - \frac{1}{a_k}$$

$$P(w_{ik} = 2) = \frac{1}{5a_k}$$

$$P(w_{ik} = 3) = \frac{1}{5a_k}$$

Where w_{ik} is the perturbation weight associated with the i^{th} household in strata k and $a_k \geq 1$.

The choice of this particular distribution is arbitrary, and other distributions can be considered. We have chosen a simple distribution for sake of illustrating the general method. The reason that the perturbation weights are not assigned symmetrically about 1 is that assigning the perturbation weights to 0 or 2 would probably not give us sufficient protection against masking uniques since uniques are assigned to be cells with a count of 1 or 2. By allowing perturbation weights to be reassigned as 3 gives us additional protection against disclosure of uniques. The probabilities assigned to each value have been chosen (a) to assign a large probability to non-perturbation (i.e. perturbation weight of one) and (b) to ensure the method of perturbation weight assignment is unbiased, i.e. the expected value of the perturbation weight is 1.

Let n_k be the number of households in strata k. The expected number of households with perturbation weights not equal to 1 is n_k/a_k in stratum k. a_k may then be set to values such as n_k or 0.5 n_k in order to achieve an expected value of 1 or 2 households per stratum with perturbation weights not equal to 1.

The expected value and the variance of the perturbation weight can be derived as follows.

$$E(w_{ik}) = 1 - \frac{1}{a_k} + \frac{2}{5a_k} + \frac{3}{5a_k}$$
$$= 1$$

$$\begin{aligned}\text{var}(w_{ik}) &= 1-\frac{1}{a_k}+\frac{4}{5a_k}+\frac{9}{5a_k}-1\\ &=\frac{8}{5a_k}\\ &=\frac{1.6}{a_k}\end{aligned}$$

In the case of census tables the perturbation weight is equivalent to the final estimation weight. All cell counts in the output tables can be written as a summation over the relevant weights associated with the contributing units. Every cell in a table represents some census subpopulation. Suppose we are interested in measuring the variance associated with c_A, the cell count associated with subpopulation A. This can be written as:

$$\begin{aligned}\text{var}(c_A) &= \text{var}(\sum_k \sum_{i\in A} w_{ik})\\ &=\sum_k \sum_{i\in A} \text{var}(w_{ik})\\ &=\sum_k \sum_{i\in A} \frac{1.6}{a_k}\\ &=1.6\sum_k \frac{m_k}{a_k}\\ &=1.6\sum_k p_k c_k\end{aligned}$$

where m_k is the total number of households in stratum k in subpopulation A,

$p_k = \frac{m_k}{n_k}$ is the proportion of stratum k in subpopulation A, and

$c_k = \frac{n_k}{a_k}$ is the expected number of households with perturbation weight not equal to 1 in stratum k.

This variance is bounded by $1.6 \Sigma c_k$. c_k can then be chosen to balance the degree of protection provided by perturbing an expected c_k records in stratum k with the information loss associated with a variance of $1.6 \Sigma c_k$, which is realised when $p_k = 1$ for all strata k.

The above method ensures consistency because all tables are produced from the same microdata and tables are therefore not rounded independently. We have protected against geographical and subpopulation differencing because we cannot be sure that an observed count of 1 or 2 is a real count of 1 or 2. A cell count of 0 also does not

guarantee that it was originally a zero and therefore the 100% cell issue is also protected against.

The above method appears promising, however the properties of the method will still need to be further investigated. It may not even be necessary to have strata defined by household size groups 1, 2, 3, 4, 5+. It may be just as efficient to have groups 1, 2, 3+. The smaller the number of household size groups the smaller the variance of c_A will be.

8. TECHNIQUE 6: IMPROVING CONSISTENCY BY ASSIGNING RANDOM NUMBERS OR RECORD KEYS TO MICRODATA FILES

Another technique for improving consistency between tables is to assign permanent random numbers to each record on the microdata file, and to use these permanent random numbers to generate random perturbations / roundings. For example, if using technique 2 – the cell perturbation technique – then assign each record on the microdata file a permanent random number between 0 and 1. Use this number to generate a value of d_{ij}, e.g. if the permanent random number is less than p_i then $d_{ij} = 0$, if between p_i and $(2+p_i)/3$ then $d_{ij} = -1$, or otherwise $d_{ij} = 2$. This technique ensures that if the same table is generated independently on more than one occasion, the same confidentialised table will be output each time. Further, any specified estimate (defined as the sum of a specified set of microdata) will be perturbed consistently within any table that includes it within one of its cells.

A simple extension of this idea uses a record key rather than a random number. For example, assign each record a key in the form of a 32-bit binary number. This record key can be used as a seed for a pseudo-random number generating function, which in turn could be used for generating d_{ij} at record level. Record keys can also be combined across records to guarantee consistent results are applied to aggregates of records. This can be done using the XOR (exclusive or) function. The XOR function will return another 32-bit binary number, and will always return the same result from the input, regardless of the order in which the individual keys are XORed together. This means any aggregate of n records will correspond to a unique 32-bit aggregate key, obtained by XORing the keys of the individual records. The key of the aggregate can be used to seed a pseudo-random number, which in turn can be used to determine if the aggregate is rounded / perturbed up or down.

For example, suppose we wanted to ensure consistency of a random rounding base 3 method. A cell value of 4 will be rounded down to a value of 3 with probability 2/3, and rounded up to a value of 6 with probability 1/3. The record keys of the four contributing records are XORed together to produce a key corresponding to the cell estimate of 4, which in turn is used as a seed to produce a pseudo-random number. If this random number is less than 2/3 the cell is rounded down to a value of 3, otherwise it is rounded up to a value of 6. Any cell holding the aggregate of the same 4 records will always be rounded in the same direction.

9. SUMMARY

This paper has illustrated several techniques that can be applied.

Tabular techniques that can be used include cell perturbation and random rounding (a special case of cell perturbation). Random rounding is a relatively common technique, is well-supported by software and is well understood. Other cell perturbation techniques allow a greater level of control over the resulting disclosure risk and information loss. Rounding and perturbation do not, in general, lead to consistent output. However the degree of consistency can be improved by using techniques such as assigning random numbers to the microdata records and using these to inform perturbation decisions, or by disaggregating tables into 'building block' tables and limiting protection action to the most vulnerable tables – those relating to the smallest subpopulations.

Microdata techniques seek to apply protections to the microdata before output is produced, therefore guaranteeing consistency and additivity in output, although this is in general achieved by introducing a greater degree of information loss, and making the information loss harder to quantify. Microdata techniques that have been canvassed are data swapping, post-randomisation method (PRAM) and assigning perturbation weights to microdata.

A final strategy may well incorporate more than one technique, e.g. data swapping in conjunction with random rounding, or different techniques for different types of output.

In evaluating techniques we need to determine the practical impacts that flow from the characteristics we have considered (or additional characteristics that we have not identified). Are some techniques more suitable for some variables than for other variables? Are some techniques more suitable for certain types of table output (e.g. small tables or high-dimension tables)? Are some techniques more suitable for certain types of analyses that users may want to conduct on the tabular output?

In evaluating a final strategy that combines a number of techniques we need to consider where techniques complement each other well, and where combining techniques can be counter-productive. A final strategy also need to take account of the computational complexity to ensure the final confidentialisation strategy can be used to give a quick turnaround for complex tables.

10. REFERENCES

Armitage, P., Merrett, K., Lyons, A. and Tame, E. (2003) *Neighbourhood Statistics in England and Wales: Disclosure Control Problems and Solutions*, Monographs of official statistics – Work session on statistical data confidentiality – Luxembourg, 7–9 April, 2003.

Australian Bureau of Statistics (2004) *Australian Standard Geographical Classification (ASGC) – Electronic Publication*, ABS cat. no. 1216.0 .

Brown, D., Lowthian, P. and Armitage, P. (2003) *Different Approaches to Disclosure Control Problems Associated with Geography*, Monographs of official statistics – Work session on statistical data confidentiality – Luxembourg, 7–9 April, 2003.

de Wolf, P.-P., Gouweleeuw, J.M., Kooiman, P. and Willenborg, L. (1998) "Reflections on PRAM", proceedings of the conference "Statistical Data Protection", 25–27 March 1998, Lisbon, Portugal. (http://neon.vb.cbs.nl/casc/related/Sdp_98_2.pdf).

de Wolf, P.-P. and Van Gelder, L. (2004) "An Empirical Evaluation of PRAM", Discussion paper No. 04012, Statistics Netherlands. (http://neon.vb.cbs.nl/casc/Related/discussion–paper–04012.pdf).

Doyle P., Lane, J. I., Theeuwes, J.J.M. and Zayatz, L.V. (eds) (2001) *Confidentiality, Disclosure, and Data Access – Theory and Practical Applications for Statistical Agencies*, Amsterdam; London: North Holland.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and de Wolf, P.-P. (1998) "Post Randomisation for Statistical Disclosure Control: Theory and Implementation", *Journal of Official Statistics*, Vol. 14, No. 4, 1998, pp. 463–478.

Lowthian, P. and Merola, G. (2004) "The application of controlled rounding for tabular data with particular reference to the Tau-Argus software", conference paper, "Methods for Statistics for UK Countries and Regions", 28 June, 2004. (http://www.statistics.gov.uk/events/downloads/SessionF2.doc).

Salazar–Gonzalez, J.J. (2002) "Controlled Rounding and Cell Perturbation: Statistical Disclosure Limitation Methods for Tabular Data". Technical paper, University of La Laguna, Tenerife, Spain (2002).

Salazar–Gonzalez, J.J., Young, C., Lowthian, P., Merola, G., Bond, S. and Brown, D. (2004) "Getting the best results in Controlled Rounding with the Least Effort". Proceedings of the CASC Project Final Conference, Barcelona. Springer–Verlag (2004).

Schubert, P. (2005) "Statistical disclosure control of microdata files accessed by remote access data laboratories", 55th Session of the International Statistical Institute (ISI), 5–12 April, 2005, contributed paper.

Shlomo, N. (2005) "Statistical Disclosure Control Methods for Census Outputs", London School of Economics and Political Science, British Society for Population Studies, Day Meeting, 11 January, 2005, Office of National Statistics presentation slides.
(http://www.lse.ac.uk/collections/BSPS/ppt/Shlomo_Disclosure_Control_Meeting.ppt)

Van Den Hout, A. and Van Der Heijden, P.G.M (2004) "The Analysis of Multivariate Misclassified Data With Special Attention to Randomized Response Data", *Sociological Methods and Research*, Vol. 32, No. 3, pp. 384–410.

Zayatz, L. (2003) *Disclosure Limitation for Census 2000 Tabular Data*, Monographs of official statistics – Work session on statistical data confidentiality – Luxembourg, 7–9 April, 2003.

FOR MORE INFORMATION . . .

INTERNET	**www.abs.gov.au** the ABS web site is the best place for data from our publications and information about the ABS.
LIBRARY	A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our web site for a list of libraries.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our web site, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE	1300 135 070
EMAIL	client.services@abs.gov.au
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS	www.abs.gov.au

2000001524213
ISBN 0 642 48176 8

RRP $11.00